日本語書籍タイトルの形式的構造の分析

矢田　竣太郎 †　　岩井　美樹 ‡　　影浦　峡 †

† 東京大学　教育学研究科

‡ 東京大学　学際情報学府

†{shuntaroy, kyo}@p.u-tokyo.ac.jp　‡mikii@g.ecc.u-tokyo.ac.jp

1　はじめに

書籍のタイトル（書名）は固有表現の一種と捉えられ、実際「関根の拡張固有表現階層」[1] においては「製品名」配下の「出版物名」の一部を成す。自動的に書名を抽出することを考えると、MUC (Message Understanding Conference) や IREX (IR and IE evaluation project in Japanese) で定義されたような伝統的な固有表現（人名、地名、組織名など）と比べ、次のような特徴を指摘できる。抽出に困難さを与えるという意味では、(i) 名詞句とは限らず、文であることすらあり、それ自身として表記上の規則性を持たないうえ、(ii) 他の固有表現を含むものが多数あることを挙げられる。一方、むしろ (i), (ii) により (iii) 書名を表すための社会的なルール（例：『』で括る）が策定されていること、及び国家等の施策により (iv) 包括的な書名リストが整備され利用可能であることは、書名抽出への手がかりをもたらす。

筆者らは書名と合致する文字列（書名文字列）を含むツイートから実際に図書に言及しているツイートを識別するというタスクに取り組んでおり、包括的な書名のリストを前提とすることで、この問題をテキスト分類タスクに落とし込んだ [2, 3]。すなわち、書名と合致する文字列を含むツイートを収集した上で、そのツイートに現れる書名文字列の周囲の文脈情報（及びその他のメタ情報）を用い、実際にその書名で表される書籍に言及しているものかそうでないかを、教師付き機械学習の手法で分類するという手法である。これまで筆者らは分類においてツイートデータに出現する書名そのものが含む情報を十分に考慮できていなかったが、上述の困難さを考慮したとしても、書名にはある種の形式的な特徴があるのではないかと考えた。

そこで本研究では、書名にいかなる形式的な特徴があるか分析するための予備的な調査として、包括的な書名リストを対象に書名の形式的構造の記述を試みる。

2　関連研究

書名の形式を分析する研究は多くはないが、これまでにいくつかの試みがなされている。影浦ら [4] はある書誌を対象に書名に使われる語彙の分析を行っているほか、海野ら [5] は書名と意味カテゴリを対応付けたうえで、そのカテゴリ間で特徴の比較分析を行っている。また山中ら [6] は日本語の書名 2,500 件程度について、文字種、形式的構造、モダリティの違いを書籍の分野と売上による差異で比較しており、本研究の関心と最も近い。

その他、前節で述べた書名の固有表現抽出上の特徴について類似がみられる他の固有表現を対象とした研究が散見される。例えば論文 [7] や映画 [8] のタイトルである。

本研究はこれらの研究の手法を参考としつつも、3 節で述べるように日本語書籍のタイトルをごく最近のものまで包括的に扱う点、あくまで意味論に立ち入らない範囲の形式的特徴に焦点を絞る点に違いがある。

3　データ

本研究では、国立情報学研究所が運営する書籍の連想検索システム Webcat Plus[1] に 2016 年 5 月までに採録された日本語書籍のタイトルのうち、ISBN を付与されたものを対象とする。Webcat Plus の書誌は、日本国内の主要な書籍データベースに収録された書誌を統合したもの[2]であり、十分な網羅性があるとみなせる。ただし分析を簡単にするため、書名よりも表現において自由度の高いサブタイトル（副書名）は含めない。また、ごく一部の書誌レコードで書誌情報の記載がコントロールされていなかった[3]ため、分析対象から

[1] http://webcatplus.nii.ac.jp/

[2] http://webcatplus.nii.ac.jp/faq_001.html#pid001

[3] たとえば、書名フィールドの中に著者名や出版社名が混在しているものなど。

外したものが 2,000 件程度ある。最終的に 1,477,278 件の書名を分析対象とした。

4 分析

4.1 方法

テキストとして記述される言語表現を文法的単位で階層化するとき、文字、単語、句、節、文の順に分けることができる。そこで大きく (1) 文字、(2) 単語、(3) 句・節・文と階層を分け、それぞれの観点で書名を分析していくことにする。ここで句以上の単位を 1 つにまとめたのは、節や文の形式の書名は存在するとはいえ、句と比べると少ないだろうと推測されるためである。もっとも、この推測の正当性も本研究では分析の対象となる。

上記の階層に基づき、本研究で分析するのは以下の項目とした。

1. 文字レベル
   (a) 文字種ごとの頻度
   (b) 書名の文字数の頻度
2. 単語レベル
   (a) 1 単語ごとの頻度
   (b) 書名の単語数の頻度
   (c) 品詞ごとの頻度
   (d) 1 書名あたりの品詞の平均頻度
3. 句・節・文レベル
   (a) 書名の文節数の頻度
   (b) 句・節・文の頻度
   (c) 頻出構文の抽出

4.2 結果

4.2.1 文字レベル

まず 1(a) 文字種ごとの頻度を算出した。ただし、ここでいう文字種とはひらがなや漢字といった区別を指し、本研究では便宜的に Unicode における分類に従った。表 1 に出現した文字種のすべてとその頻度を示す。ここからは、書名を構成する文字の大部分が漢字であり、カタカナよりもひらがなの方が多く使われていることがわかる。

次に 1(b) 書名ごとに文字数をカウントした。表 2 にその要約統計量を示す。また文字数ごとの頻度について、度数分布表のピーク部分を図 1 に示した。以上からは、書名は概ね 11 文字前後のものが多いことがわかる。

表 1: 文字種ごとの頻度

文字種	頻度	割合
漢字	7,419,766	42.8%
ひらがな	3,914,386	22.6%
カタカナ	3,141,742	18.1%
アルファベット	1,457,518	8.4%
記号	806,650	4.7%
数字	301,568	1.7%
濁点・半濁点・長音記号	290,278	1.7%
合成漢字・部首漢字など	4,927	-
その他の言語の文字	1,824	-

表 2: 書名長の要約統計量（文字数・単語数・文節数）

要約統計値	文字数	単語数	文節数
平均	11.7	5.7	2.2
標準偏差	7.3	3.1	1.2
最小値	1	1	1
第 1 四分位点	7	4	1
中央値	10	5	2
第 3 四分位点	14	7	3
最大値	248	78	29

4.2.2 単語レベル

単語レベルの分析結果を行うにあたり、単語は形態素とし、形態素解析器 MeCab[4] [5]でそのように判定されたものを用いる。まず 2(a) 1 単語ごとの頻度について、上位 10 件を表 3 に示した。一方、表 4 は内容語のみで計数した頻度に基づく上位 10 件の単語である。このとき、活用する単語については基本形（原形）を用い、表層形が同じでも品詞が違えば区別した。ただし IPA 品詞体系に基づく。

次に、2(b) 書名の単語数について表 1 に示した。度数分布は文字数の場合と同様の傾向を示したため、省略する。この結果から、単語数 6 前後の書名が多いことがわかる。

2(c) 品詞ごとの頻度についても調査したところ、「名詞-一般」が最も多く、全体の約 34%を占めていた。次

[4] http://taku910.github.io/mecab/
[5] 辞書は一般的に使われる IPADIC とした。

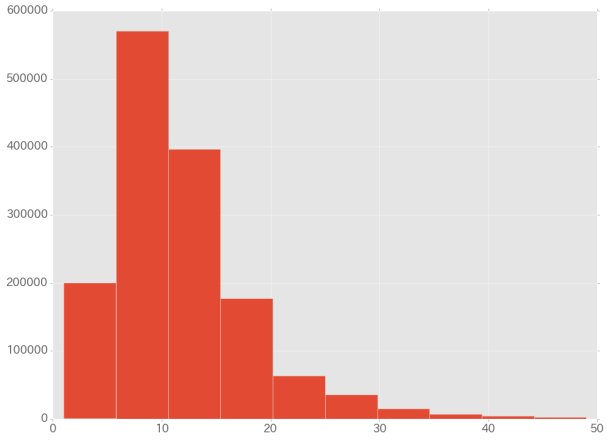

図 1: 書名文字数の度数分布（ピーク部）

表 3: 1 単語ごとの頻度（上位 10 件）

単語（品詞）	頻度	累積割合
の [助詞-連体化]	665,737	7.88%
・ [記号-一般]	151,734	9.68%
と [助詞-並立助詞]	102,751	10.89%
を [助詞-格助詞]	101,938	12.10%
！ [記号-一般]	82,090	13.07%
に [助詞-格助詞]	80,366	14.02%
で [助詞-格助詞]	65,060	14.79%
が [助詞-格助詞]	62,113	15.53%
「 [記号-括弧開]	57,091	16.21%
」 [記号-括弧閉]	57,065	16.88%

点の「名詞-サ変接続」はその 3 分の 1 以下 (10%) であり、「名詞」が大部分を占める (66%) ことがわかった。

最後に 2(d) 1 書名あたりの品詞の平均頻度を調べたところ、書名は約 4 個の名詞と約 1 個の助詞を平均して含むといえた。全品詞の平均頻度値は割愛する。

4.2.3 句・節・文レベル

句・節・文レベルの分析の出発点として、3(a) 書名を分節に分けた場合の長さを測定する。文節の判定には係り受け解析器の CaboCha[6] を用いた。1(b), 2(b) と同様に、表 1 に要約統計量を示した。度数分布は文字数、単語数の場合と同様の傾向を示したため、省略する。これより、書名は 2 文節前後のシンプルな構造のものが中心であるとわかる。

3(b) 書名が句・節・文のいずれであるかをカウントする。このとき日本語文法上の規則に基づき、次のように判定した。

表 4: 1 単語ごとの頻度（上位 10 件; 内容語のみ）

単語（品詞）	頻度	累積割合
する [動詞-自立]	36,166	0.57%
日本 [名詞-固有名詞]	33,379	1.10%
集 [名詞-接尾]	30,660	1.58%
学 [名詞-接尾]	29,581	2.05%
ため [名詞-非自立]	27,079	2.47%
法 [名詞-接尾]	25,128	2.87%
わかる [動詞-自立]	23,763	3.24%
入門 [名詞-サ変接続]	23,442	3.61%
問題 [名詞-ナイ形容詞語幹]	20,623	3.94%
本 [名詞-一般]	19,853	4.25%

句 最後の文節が名詞で終わるもの

文 感動詞を含むもの（独立語文）、最後の文節が終助詞で終わるもの、最後の文節に述語（動詞・形容詞・形容動詞）を含み基本形（終止形）であるもの

節 句と文以外のもの

この場合、節には倒置形の文も含んでしまうが、動詞による連体修飾との区別が煩雑であるため、許容することとした。計数した結果、句が 87.9%、文が 7.5%、節が 4.6%となり、基本的には書名は句（名詞句）であることが確認された。

最後に、3(c) 書名に頻出の構文の抽出を試みる。ここで構文とは、内容語を品詞名に置き換え、機能語を表層形のままに保つことで表現できる文法的な構造をいうものとする（例：“[名詞] の [名詞]”）。ただし、我々は 2(a) の分析により書名に頻出の単語を得ていることから、出現頻度（内容語のみ）の累積割合で 10%以内の内容語も表層系をそのまま用いることで、より書名らしい構文の抽出が可能であると着想した。さらに、名詞の連続を [名詞句] と抽象化して、より大局的な構文を俯瞰すると、表 5 を得る。

この結果からは、上位 2 件の構文を合わせても 3 割未満で、残り 7 割は多種多様なパターンがロングテール状に分布していることがわかる。しかしながら、どのパターンにおいても書名は名詞句を中心としたあまり複雑すぎない構文を用いていると総論できる。

[6] https://taku910.github.io/cabocha/

表 5: 頻出構文（上位 20 件）

構文パターン	頻度	割合
[名詞句]	243,144	16.46%
[名詞句] の [名詞句]	156,083	10.57%
[名詞句] と [名詞句]	24,336	1.65%
[名詞句]・[名詞句]	19,714	1.33%
[名詞句] の [名詞句] と [名詞句]	11,541	0.78%
[名詞句] 集	8,307	0.56%
[名詞句] の [名詞句] の [名詞句]	8,292	0.56%
[名詞句] と [名詞句] の [名詞句]	8,076	0.55%
[名詞句]・[名詞句] の [名詞句]	6,037	0.41%
[形容詞][名詞句]	5,989	0.41%
[動詞][名詞句]	5,702	0.39%
[名詞句] を [動詞]	5,687	0.38%
[名詞句] 入門	5,513	0.37%
[接頭詞][名詞句]	4,531	0.31%
[名詞句] を [動詞][名詞句]	4,511	0.31%
[名詞句] への [名詞句]	4,464	0.30%
[名詞句]・[名詞句]・[名詞句]	4,287	0.29%
[名詞句][接頭詞][名詞句]	4,133	0.28%
[名詞句] & [名詞句]	4,100	0.28%
[名詞句] 論	4,071	0.28%

5 課題と展望

本研究では、日本語書籍のタイトルに用いられる言語表現の形式的構造に焦点を当てて分析したが、分野（ジャンル）やクラス（新書、文庫の区別など）といった書籍の社会的な配置・属性は考慮していない。本研究の分析項目について書名リストを出版社やジャンルで分けたり、出版年をもとに経年の傾向を見たりすることで顕になる特徴があるはずであり、こうした書名の外側の情報と関連付けることは今後の課題としたい。

また、固有表現としての書名の特徴を知るためには当然、他の固有表現クラスとの比較が重要である。書名に特徴が類似する固有表現には映画タイトルや楽曲タイトルといった芸術作品名が該当するが、書籍の世界ではメディアミックス施策による映画等の書籍化（あるいはその逆）や書籍としての楽譜集の存在があり、それらを形式の上で区別できるかどうかも関心がある。

謝辞

本研究は科学研究費補助金挑戦的萌芽研究「オンラインを介して「前読書家」の読書を触発する方式・環境の開発」（課題番号 16K12542）の支援を得て行われた。また、本研究で用いた Webcat Plus の書誌データは，国立情報学研究所特任准教授 阿辺川武氏に提供を受けた。

参考文献

[1] Sekine, Satoshi and Chikashi Nobata (2004) “Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy,” LREC2004, Lisbon, Portugal.

[2] Yada, Shuntaro and Kyo Kageura (2016) “Improved Identification of Tweets that Mention Books: Selection of Effective Features,” ICADL2016, Tsukuba, Japan.

[3] Yada, Shuntaro and Kyo Kageura (2015) “Identification of Tweets that Mention Books: An Experimental Comparison of Machine Learning Methods,” ICADL2015, Seoul, Korea.

[4] 影浦峡・海野敏・戸田慎一 (1988)「書誌の書名の語彙分析」,『書誌索引展望』, 第 12 巻, 第 2 号, pp. 1–16.

[5] 海野敏・影浦峡・戸田慎一 (1989)「書誌の書名構成語の数量的分析: 意味カテゴリーの共出現状況」,『図書館学会年報』, 第 35 巻, 第 3 号, pp. 116–125.

[6] 山中信彦・毛愛涛 (2015)「書名の言語学 : 国立国会図書館サーチに基づいて」,『計量国語学』, 第 30 巻, 第 1 号, pp. 14–31.

[7] 里見香奈・成田健一. (2016) 「『自己』にかかわる心理学的研究の計量書誌学的分析 : わが国の学会誌に掲載された実証論文のタイトル分析 : 1980 年-2013 年」,『関西学院大学心理科学研究』, 第 42 巻, pp. 25–32.

[8] 小澤真冬 (2015) 「キネマ旬報ベスト・テンに見る映画のタイトル分析」,『語文』, 第 151 巻, pp. 126–111.